Developing Optimal Search Strategies for Detecting Clinically Sound Causation Studies in MEDLINE

Nancy L. Wilczynski, MSc, R. Brian Haynes, MD, PhD, for the Hedges Team
Health Information Research Unit, McMaster University, Hamilton, Ontario, Canada

Abstract

Background: Clinical end users of MEDLINE must be able to retrieve articles that are both scientifically sound and directly relevant to clinical practice. The use of methodologic search filters has been advocated to improve the accuracy of searching for such studies. These filters are available for the literature on therapy and diagnosis, but strategies for the literature on causation have been less well studied.
Objective: To determine the retrieval characteristics of methodologic terms in MEDLINE for identifying methodologically sound studies on causation.
Design: Comparison of methodologic search terms and phrases for the retrieval of citations in MEDLINE with a manual hand search of the literature (the gold standard) for 161 core health care journals.
Methods: 6 trained, experienced research assistants read all issues of 161 journals for the publishing year 2000. Each article was rated using purpose and quality indicators and categorized into clinically relevant original studies, review articles, general papers, or case reports. The original and review articles were then categorized as 'pass' or 'fail' for methodologic rigor in the areas of therapy/quality improvement, diagnosis, prognosis, causation, economics, clinical prediction, and review articles. Search strategies were developed for all categories including causation.
Main outcome measures: Sensitivity, specificity, precision, and accuracy of the search strategies.
Results: 12% of studies classified as causation met basic criteria for scientific merit for testing clinical applications. Combinations of terms reached peak sensitivities of 93%. Compared with the best single term, multiple terms increased sensitivity for sound studies by 15.5% (absolute increase), but with some loss of specificity when sensitivity was maximized. Combining terms to optimize sensitivity and specificity achieved sensitivities and specificities both above 80%.
Conclusion: The retrieval of causation studies cited in MEDLINE can be substantially enhanced by selected combinations of indexing terms and textwords.

Introduction

With the increasing emphasis on evidence-based medicine, clinicians must be able to identify the best evidence to plan effective patient care. This task is difficult because advances in health care practice are published in a wide array of journals, mixed with many preliminary studies. Online searching for best evidence through electronic databases such as MEDLINE results in the user searching through approximately 5,000 journals with an estimated 8,000 citations entered on a weekly basis. This explosion of information makes it difficult for clinicians to keep up to date with advances in health care [1, 2] and as a result most researchable information needs are unmet [3]. Even clinicians who support evidence-based medicine in principle often believe they do not do this in practice [4]. When they do try to find research evidence, practitioners do not search the medical literature very effectively [5]. If large electronic bibliographic databases are to be helpful to clinicians, they must be able to retrieve articles that are scientifically sound and directly relevant to the health problem they are trying to solve, without missing key studies, or retrieving excessive numbers of irrelevant or misleading studies.

One possible aid is to develop methodologic search filters to improve the retrieval of clinically relevant and scientifically sound study reports from large biomedical research bibliographic databases, such as MEDLINE. In MEDLINE, filters are created by adding, to the usual disease content terms, Medical Subject Headings (MeSH), explosions (px), publication types (pt), subheadings (sh) and textwords (tw) that detect research design features indicating methodologic rigor for applied health care research, for instance, 'Exp myocardial infarction and (randomized controlled trial (pt) or clinical trial (pt))'. The use of such methodologic search filters has been advocated [6], and filters have been developed to improve the accuracy of searching for such studies [7, 8, 9]. Most of the studies have focused on information retrieval for therapy and diagnostic articles as well as systematic reviews. Little work has been done in the area of causation.

In the early 1990s, our group developed search filters on a small subset of 10 journals and for 4 types of journal articles (therapy, diagnosis, prognosis and causation [etiology]) [10, 11], and these strategies have been adapted for use in the Clinical Queries interface of MEDLINE (http://www.ncbi.nlm.nih.gov/entrez/query/static/clinical.html). This research is being updated and expanded using data from 161 journals from the publishing year 2000. The robustness of empirical search strategies developed in 1991 for detecting clinical content in MEDLINE in the year 2000 has already been reported [12]. In this paper, we report on the information retrieval properties of single terms and combinations of terms in MEDLINE for identifying methodologically sound studies on the causation of disorders due to the magnitude on data generated from this study.

Methods

The study compared the retrieval performance of methodologic search terms and phrases in MEDLINE with a manual review of each article for each issue of 161 journal titles for the year 2000. MeSH terms and textwords related to research design features of studies of causation were run as search strategies. These search strategies were treated as "diagnostic tests" for sound studies and the manual review of the literature was treated as the "gold standard." The sensitivity, specificity, precision, and accuracy of MEDLINE searches were determined. Sensitivity for a given topic is defined as the proportion of high quality articles for that topic that are retrieved; specificity is the proportion of low quality articles not retrieved; precision is the proportion of retrieved articles that are of high quality; and accuracy is the proportion of all articles that are correctly classified. Sensitivity and specificity are not affected by the proportion of high quality articles in the database; precision is dependent on this proportion, and so is accuracy to a lesser extent.

Six research assistants reviewed 161 journals titles for the year 2000, and applied methodologic criteria to each item in each issue to determine if the article was methodologically sound for 7 purpose categories (two other types of articles, cost and qualitative studies, were also classified but had no rigor criteria). All purpose category definitions and corresponding methodologic rigor were outlined in a previous paper [13]. The methodologic criteria applied for studies of causation are shown in Table 1.

The 161 journal titles reviewed in 2000 were chosen in an iterative process based on recommendations of clinicians and librarians, Science Citation Index Impact Factors, and ongoing assessment of their yield of studies and reviews of scientific merit and clinical relevance for the disciplines of internal medicine, general medical practice, mental health, and general nursing practice (list of journals provided by the authors upon request). Research staff were rigorously calibrated and inter-rater agreement for application of methodologic criteria exceeded 80% beyond chance for all study purpose categories [13].

To construct a comprehensive set of search terms, we began a list of MeSH terms and textwords and then sought input from clinicians and librarians in the United States and Canada through interviews of known searchers, requests at meetings and conferences, and requests to the National Library of Medicine. Individuals were asked what terms or phrases they used when searching for studies of causation, prognosis, diagnosis, treatment, economics, clinical prediction guides, reviews, costs, and of a qualitative nature. Terms could be from MeSH, including publication types (pt), check tags, and subheadings (sh), or could be textwords (tw) denoting methodology in titles and abstracts of articles. We compiled a list of 5,345 terms (list of terms tested provided by the authors upon request). The database was randomly split components of 60% and 40%. Search strategies were initially tested and developed in 60% of the database (development) and then validated in 40% of the database (validation).

Results

49,028 articles were identified after matching the hand search records with the data downloaded from MEDLINE. Of these 2,421 were classified as causation, of which 282 (12%) were methodologically sound. Table 2 shows the best single terms for high-sensitivity, high-specificity, and best balance of sensitivity and specificity from the development database and the operating characteristics of these terms in the validation database. Small absolute differences were noted when comparing the performance in the development and validation databases. In most cases terms

Table 1 – Methodologic Rigor Applied for Studies of Causation

Purpose Category	Methodologic Rigor
Causation	Observation concerned with the relationship between exposures and putative clinical outcomes; Data collection was prospective; Clearly identified comparison group(s); Blinding of observers of outcome to exposure.

Table 2 – Single Terms with the Best Sensitivity (keeping Specificity ≥50%), Best Specificity (keeping Sensitivity ≥50%), and Best Optimization of Sensitivity and Specificity (based on abs[sensitivity-specificity]<10%) for Detecting Studies of Causation in MEDLINE in 2000

Search term - OVID search	Sensitivity (%) Development Validation Difference (95% CI)*	Specificity (%) Development Validation Difference (95% CI)*	Precision (%) Development Validation Difference (95% CI)*	Accuracy (%) Development Validation Difference (95% CI)*
Best Sensitivity Risk:.mp.	77.6 75.9 -1.7 (-12.2 to 8.2)	83.5 83.2 -0.3 (-1.0 to 0.4)	2.7 2.4 -0.3 (-1.0 to 0.4)	83.4 83.1 -0.3 (-1.0 to 0.4)
Best Specificity Risk factor:.mp.	50.6 49.1 -1.5 (-13.4 to 10.5)	90.8 90.8 0.0	3.2 2.9 -0.3 (-1.3 to 0.8)	90.5 90.6 0.1 (-0.5 to 0.6)
Best Optimization of Sensitivity & Specificity Risk:.mp.	77.6 75.9 -1.7 (-12.2 to 8.2)	83.5 83.2 -0.3 (-1.0 to 0.4)	2.7 2.4 -0.3 (-1.0 to 0.4)	83.4 83.1 -0.3 (-1.0 to 0.4)

*Comparing the development and validation data sets. Differences are not statistically significant.

Table 3 – Combination of Terms with the Best Sensitivity (keeping Specificity ≥50%), Best Specificity (keeping Sensitivity ≥50%), and Best Optimization of Sensitivity and Specificity (based on abs[sensitivity-specificity]<1%) for Detecting Studies of Causation in MEDLINE in 2000

Search strategy - OVID search	Sensitivity (%) Development Validation Diff (95% CI)*	Specificity (%) Development Validation Diff (95% CI)*	Precision (%) Development Validation Diff (95% CI)*	Accuracy (%) Development Validation Diff (95% CI)*†
Best Sensitivity Risk:.mp. OR Exp cohort studies OR Between group:.tw.	93.1 93.5 0.4 (-6.5 to 6.3)	63.2 63.1 -0.1 (-1.0 to 0.8)	1.5 1.4 -0.1 (-0.4 to 0.6)	63.4 63.3 -0.1 (-1.0 to 0.8)
Best Sensitivity - Small decrease in Sensitivity with large increase in Specificity Risk:.mp. OR Exp cohort studies OR Mortality.tw.	90.8 86.1 -4.7 (-13.3 to 2.8)	72.3 72.0 -0.3 (-1.0 to 0.5)	1.9 1.7 -0.2 (-0.7 to 0.2)	72.4 72.1 -0.3 (-1.1 to 0.5)
Best Specificity Relative risk:.tw. OR Risks.tw. OR Cohort stud:.mp.	51.2 52.8 1.6 (-10.3 to 13.5)	94.8 94.9 0.1 (-0.4 to 0.4)	5.6 5.5 -0.1 (-1.9 to 1.7)	94.6 94.7 0.1 (-0.3 to 0.5)
Best Specificity - Small decrease in Specificity with large increase in Sensitivity Cohort:.tw. OR Confidence interval:.tw. OR Relative risk:.tw.	60.3 61.1 0.8 (-11.0 to 12.1)	92.5 92.6 0.1 (-0.4 to 0.6)	4.5 4.4 -0.1 (-1.5 to 1.2)	92.3 92.4 0.1 (-0.3 to 0.6)
Best Optimization of Sensitivity & Specificity Risk.mp. OR Mortality.mp. OR Cohort.tw.	83.3 80.6 -2.7 (-12.6 to 6.2)	82.9 82.8 -0.1 (-0.8 to 0.6)	2.8 2.5 -0.3 (-1.0 to 0.4)	82.9 82.8 -0.1 (-0.8 to 0.6)

*Comparing the development and validation data sets. Differences are not statistically significant.

performed a little less well in the validation database, but the maximal difference was only 1.7%. When maximizing sensitivity while keeping specificity ≥50%, sensitivities >75% were achieved. When maximizing specificity while keeping sensitivity ≥50%, specificities >90% were achieved but this occurred at the expense of sensitivity. When optimizing sensitivity and specificity the best single term was the same as that reported for best sensitivity (Table 2).

Three term strategies are shown in Table 3. As expected, combinations increased sensitivity. Sensitivities of >90% can be achieved when combining terms with specificity remaining at 72%. Once again the results were trivially different when comparing the performance between the development and validation databases.

Discussion

Our study documents search strategies that can help discriminate relevant from nonrelevant articles on the causation of health disorders. Those interested in all articles on causation, with time to sort out irrelevant articles, will be best served by the most sensitive search. Those with little time on their hands who are looking for a few good articles on causation will likely be best served by the most specific strategies. The strategies that optimized sensitivity and specificity while minimizing the difference between the two provide the best separation of hits from false drops but do so without regard for whether sensitivity or specificity is affected.

In all cases precision was low. This is the inevitable result of a low proportion of relevant studies for a given purpose in a very large, multipurpose database. This means that searchers will continue to need to invest their time in discarding irrelevant retrievals. While low precision in searching can be of concern, the low values here should not be over-interpreted: we did not limit the searches by clinical content terms, as would be the usual case in clinical searches. Precision can be enhanced by combining search strategies in these tables with methodologic terms using the Boolean 'AND NOT' and/or by combining search strategies with content specific terms using the Boolean 'AND".

The next phases of our project will focus on finding better search strategies through using more sophisticated strategies as outlined above.

Table 4 – Comparison of Combination of Terms with the Best Sensitivity (keeping Specificity ≥50%) and the Best Specificity (keeping Sensitivity ≥50% in 2000) for Detecting Studies of Causation in MEDLINE in 1991 and 2000

Search Strategies	Sensitivity (%) 1991 2000	Specificity (%) 1991 2000
Best Sensitivity		
1991†		
"Cohort studies" [MESH] OR "Risk" [MESH] OR ("Odds" [WORD] AND "ratio*" [WORD]) OR ("Relative" [WORD] AND "risk" [WORD]) OR "Case" control*" [WORD] OR Case-control studies [MESH]	81.7	70.2
2000‡		
Risk:.mp. OR Exp cohort studies OR Between group:.tw.	93.1	63.2
Best Specificity		
1991†		
"Case-control studies" [MH:NOEXP] OR "Cohort studies" [MH:NOEXP]	40.1	96.5
2000‡		
Relative risk:.tw. OR Risks.tw. OR Cohort stud:.mp.	51.2	94.8

†PubMed search strategy. ‡OVID search strategy.

Compared with the performance of search terms for causation that we developed in 1991, the combinations of terms in 2000 slightly out-performed 1991 strategies, but with trade-offs (Table 4). For example, for the most sensitive strategy, sensitivity rose by 11.4% but specificity fell by 7.0%.

Conclusion

Selected combinations of indexing terms and textwords can substantially enhance the retrieval of causation studies cited in MEDLINE.

References

[1] Haynes RB, Sackett DL, Tugwell P. Problems in the handling of clinical and research evidence by medical practitioners. *Arch Intern Med* 1983;143:1971-5.

[2] Martinez JL, Licea Serrato Jde D, Jimenez R, Grimes RM. HIV/AIDS practice patterns, knowledge, and education needs among Hispanic clinicians in Texas, USA, and Nuevo Leon, Mexico. *Rev Panam Salud Publica* 1998;4:14-9.

[3] Covell DG, Uman GC, Manning PR. Information needs in office practice: are they being met? *Ann Intern Med* 1985;103:596-9.

[4] Tomlin Z, Humphrey C, Rogers S. General practitioners' perceptions of effective health care. *BMJ* 1999;318:1532-5.

[5] Balas EA, Stockham MG, Mitchell JA, Sievert ME, Ewigman BG, Boren SA. In search of controlled evidence for health care quality improvement. *J Med Syst* 1997;21:21-32.

[6] Haynes RB, McKibbon KA, Fitzgerald D, Guyatt GH, Walker CJ, Sackett DL. How to keep up with the medical literature: V. Access by personal computer to the medical literature. *Ann Intern Med* 1986;105:810-6.

[7] Nwosu CR, Khan KS, Chien PF. A two-term MEDLINE search strategy for identifying randomized trials in obstetrics and gynecology. *Obstet Gynecol* 1998;91:618-22.

[8] Robinson KA, Dickersin K. Development of a highly sensitive search strategy for the retrieval of reports of controlled trials using PubMed. *Int J Epidemiol* 2002;31:150-3.

[9] Bachmann LM, Coray R, Estermann P, Ter Riet G. Identifying diagnostic studies in MEDLINE: reducing the number needed to read. *J Am Med Inform Assoc* 2002;9:653-8.

[10] Wilczynski NL, Walker CJ, McKibbon KA, Haynes RB. Assessment of methodologic search filters in MEDLINE. *Proc Annu Symp Comput Appl Med Care* 1993;:601-5.

[11] Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. *J Am Med Inform Assoc* 1994;1:447-58.

[12] Wilczynski NL, Haynes RB; Hedges Team. Robustness of empirical search strategies for clinical content in MEDLINE. *Proc AMIA Symp* 2002;:904-8.

[13] Wilczynski NL, McKibbon KA, Haynes RB. Enhancing retrieval of best evidence for health care from bibliographic databases: calibration of the hand search of the literature. *Medinfo* 2001;10:390-3.

Acknowledgments

This research was funded by the National Library of Medicine, USA. The Hedges Team includes Angela Eady, Brian Haynes, Susan Marks, Ann McKibbon, Doug Morgan, Cindy Walker-Dilks, Stephen Walter, Nancy Wilczynski, and Sharon Wong.